UCRL-JRNL-216004

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

A Comprehensive Catalog of Human KRAB-associated Zinc Finger Genes: Insights into the Evolutionary History of a Large Family of Transcriptional Repressors

S. Huntley, D. M. Baggott, A. T. Hamilton, M. Tran-Gyamfi, S. Yang, J. Kim, L. Gordon, E. Branscomb, L. Stubbs

October 7, 2005

Genome Research

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors

Stuart Huntley[1], Daniel M. Baggott[1], Aaron T. Hamilton[1], Mary Tran-Gyamfi[1], Shan Yang[1], Joomyeong Kim[2], Laurie Gordon[1], Elbert Branscomb[3], and Lisa Stubbs[1,4]

[1]Genome Biology and [3]Microbial Systems Divisions, Biosciences, Lawrence Livermore National Laboratory
7000 East Avenue, L-441, Livermore, CA 94550

[2] Present address: Louisiana State University, Baton Rouge LA

[4]To whom correspondence should be addressed, stubbs5@llnl.gov; tel. 925-422-8473; fax. 925-422-2099

Running title: Human KRAB zinc finger family

Keywords: zinc finger, gene annotation, annotation, transcription factor, gene duplication

ABSTRACT

Krüppel-type zinc finger (ZNF) motifs are prevalent components of transcription factor proteins in all eukaryotic species. In mammals, most ZNF proteins comprise a single class of transcriptional repressors in which a chromatin interaction domain, called the *Krüppel*-associated box (KRAB) is attached to a tandem array of DNA-binding zinc-finger motifs. KRAB-ZNF loci are specific to tetrapod vertebrates, but have expanded dramatically in numbers through repeated rounds of segmental duplication to create a gene family with hundreds of members in mammals. To define the full repertoire of human KRAB-ZNF proteins, we searched the human genome for key motifs and used them to construct and manually curate gene models. The resulting KRAB-ZNF gene catalog includes 326 known genes, 243 of which were structurally corrected by manual annotation, and 97 novel KRAB-ZNF genes; this single family therefore comprises 20% of all predicted human transcription factor genes. Many of the genes are alternatively spliced, yielding a total of 743 distinct predicted proteins. Although many human KRAB-ZNF genes are conserved in mammals, at least 136 and potentially more than 200 genes of this type are primate-specific including many recent segmental duplicates. KRAB-ZNF genes are active in a wide variety of human tissues suggesting roles in many key biological processes, but most member genes remain completely uncharacterized. Because of their sheer numbers, wide-ranging tissue-specific expression patterns, and remarkable evolutionary divergence we predict that KRAB-ZNF transcription factors have played critical roles in crafting many aspects of human biology, including both deeply conserved and primate-specific traits.

INTRODUCTION

The human genome is predicted to contain approximately 30,000 genes (Lander et al. 2001; Venter et al. 2001) including at least 2,000 loci that encode transcription factor proteins (TFs) (Messina et al. 2004). TFs are typically comprised of DNA binding domains, which direct the attachment of the TF to DNA at specific recognition sites, and one or more effector domains, which confer gene regulatory activities to the proteins. Generally, these effectors mediate protein-protein interactions, allowing TFs to associate with other TFs, cofactors, or chromatin-altering enzymes. The C2H2*,* or *Krüppel*-type zinc finger (ZNF), is the most common motif in eukaryotic TF DNA binding domains. The C2H2 motif consists of 28 amino acid residues with the pattern C-X2-C-X12-H-X3-H; proper folding of the peptide motif is dependent on the tetrahedral caging of a zinc ion by the paired cysteine and histidine residues. The DNA binding domains of ZNF proteins are typically comprised of multiple C2H2 motifs joined together through a conserved 7 amino acid linker sequence in tandem arrays. Proteins of this type are ancient and numerous, encoded by large and diverse gene families in all eukaryotic genomes. Subfamilies of distinct structure and function have arisen in different evolutionary lineages, defined by the types of effector domains included in the protein (Bellefroid et al. 1991; Chung et al. 2002; Knochel et al. 1989); reviewed by (Collins et al. 2001; Huntley et al. In press).

The majority of mammalian C2H2 ZNF proteins include an effector motif called the *Krüppel*-associated box, or KRAB. The KRAB domain acts as a specific binding partner for KAP-1 (product of the *TRIM28* gene), a co-repressor protein that serves to recruit chromatin-condensing histone deacetylase complexes to localized regions surrounding the DNA binding sites (Ayyanathan et al. 2003). KRAB-associated ZNF (KRAB-ZNF) proteins therefore function as potent repressors of target gene expression (Ayyanathan et al. 2003; Friedman et al. 1996; Margolin et al. 1994) Proteins combining KRAB and ZNF domains are specific to tetrapod vertebrates, but since the introduction of this gene structure the family has expanded dramatically to include hundreds of members in mammals (Bellefroid et al. 1995; Hamilton et al.

2003; Looman et al. 2002). Most KRAB-ZNF loci reside in familial gene clusters indicating that the family has evolved primarily through tandem *in situ* duplication (Bellefroid et al. 1995). Mammalian KRAB-ZNF gene clusters frequently contain scores of closely related genes, and typically include significant numbers of lineage-specific loci due to ongoing rounds of gene duplication as well as gene loss (Shannon et al. 2003). Once duplicated, the new gene copies diversify through structural changes in the zinc finger arrays that, in some cases at least, have been shown to be driven by positive selection (Hamilton et al. 2003; Schmidt and Durrett 2004; Shannon et al. 2003). Since even subtle alterations in zinc-finger array structure can yield proteins with distinct gene targets and DNA recognition specificities (Krebs et al. 2005), this striking pattern of divergence suggests an active selection for novel transcription factor proteins with altered DNA binding properties.

Although KRAB-ZNF genes comprise a significant fraction of the human transcription factor gene repertoire, most family members remain completely uncharacterized. Many genes are classified as hypothetical, and our preliminary scans of the human genome indicated that a significant number of potential ZNF protein-coding genes had not been annotated. To derive a complete catalog of this large family of human transcription factor genes, we computationally analyzed and manually curated all significant segments of the human genome containing adjacent same-strand KRAB and ZNF domains. These efforts revealed 423 complete KRAB-ZNF genes including many that produce alternative transcripts, and predict the existence of at least 743 distinct proteins. We also identified 341 pseudogene sequences, most of which correspond to partial duplication products. Analyses of genomic distribution, expression patterns, paralog relationships, and a preliminary comparison to predicted genes in other sequenced mammalian genomes permitted a genome-wide assessment of the major mechanisms through which this large gene family has evolved.

RESULTS

Assembling and Annotating the Human KRAB-ZNF Catalog

Most KRAB-ZNF genes exist as simple modular structures with effector domains and zinc-finger motifs encoded in distinct 5′ and 3′ exons, respectively (Fig. 1). These loci typically include one exon that encodes a KRAB-A motif, which plays a dominant role in KAP-1-mediated repression (Friedman et al. 1996), and a single exon encoding tandem arrays of zinc fingers; many genes also encode additional modulating motifs on separate exons. These may include KRAB-B (Bellefroid et al. 1991), a novel KRAB-B variant we will refer to as KRAB-BL (KRAB-BL exons are 30 bp larger than KRAB-B exons, extending in the 3' direction), KRAB-b (Mark et al. 1999), or KRAB-C (Looman et al. 2004).

In addition to the KRAB domains, a small but significant number of KRAB-ZNF genes encode a second effector, called SCAN, which is also vertebrate-specific (Sander et al. 2003). Many genes containing both SCAN and KRAB domains (SCAN-KRAB-ZNF genes) are clustered with KRAB-ZNF and SCAN-ZNF genes and in many clusters the three types of genes show clear evolutionary relationships (Sander et al., 2003; and see below). In some of these cases at least, the related human SCAN-ZNF or KRAB-ZNF genes appear to have been derived from ancestral SCAN-KRAB-ZNF loci that subsequently lost one of the two effector domains (Kim et al. 2004). To catalog the complete gene family, we employed profile hidden Markov model (HMM) software to identify KRAB, SCAN, and *Krüppel*-type finger motifs (HMMER 2.3, http://hmmer.wustl.edu/). Profile HMMs are statistical descriptions of sequence consensus and provide a basis for sensitively identifying additional related sequences. Profile HMMs were generated using alignments of motif peptide sequences collected from available known human genes and preliminary pattern-based searches of human genome sequences.

Based on RNA evidence and our HMMER-identified motifs, we generated models for all well-supported alternative transcripts arising from each locus (example shown in Fig. 1). We also included alternate transcripts for many previously identified genes based on RNA evidence.

We then compared our models to publicly available gene models to identify overlap with previously described human genes. For the 326 previously known loci, manual annotation produced 658 transcript models. These included 491 models that overlap transcript models in the public databases but were extended or corrected by manual annotation, and 159 transcript models from the public datasets that were not modified. In addition to the known genes, we identified 97 KRAB-ZNF loci capable of encoding full-length proteins (defined as open reading frame translations that include at least one N-terminal effector domain and an array of 2 or more zinc fingers) that are not described in the public databases. For 77 of these novel loci we were able to generate gene models capable of encoding complete KRAB-ZNF proteins, supported by public mRNA and / or spliced EST evidence. For 17 of the remaining 20 loci, the only available RNA evidence consists of unspliced EST sequences, and for three loci no RNA evidence was available. For these genes, we used the motif coordinates and additional genome features (see Methods) to generate putative gene models. Results of these gene curation efforts are summarized in Table 1; a complete list of curated genes and transcripts is provided in Suppl. Table S1. Curated mRNA sequences, predicted proteins, exon structure and other information, including links to external information sources and information on other types of *Krüppel*-type human proteins, can be accessed from the project website (http://znf.llnl.gov).

Altogether we identified 423 loci capable of encoding full-length proteins that contain both effector and zinc-finger domains. These include 30 SCAN-ZNF, 28 SCAN-KRAB-ZNF and 365 KRAB-ZNF genes (Table 1). For simplicity we will hereafter refer to the collection of genes simply as KRAB-ZNF genes, except to discuss properties specific to one of the three subtypes. In addition to these 423 loci with at least one full-length transcript, we also identified 128 genes with non-canonical structures; for example, loci predicted to encode ZNF-only, KRAB-only or SCAN-only proteins. These loci were curated as potential protein coding genes only if supported by mRNA evidence, but a number of genes of this type correspond to previously described and named loci, and therefore all are described as known or putative genes on our website (http://znf.llnl.gov). Some of these loci may encode functional genes. However, in the following discussions we will focus on the 423 loci capable of encoding full-length proteins with both effector (SCAN, KRAB or both) and zinc-finger domains.

Gene Models and Alternate Splicing in KRAB-ZNF Loci

In addition to the archetypically structured KRAB-ZNF transcripts, alternate transcripts encoding KRAB-only or fingers-only proteins with known or predicted functions have been described for several genes (Bellefroid et al. 1993; Oh et al. 2005; Wu et al. 2003). Furthermore, KRAB-ZNF loci with SCAN or multiple distinct KRAB exons are known to encode protein isoforms that include different combinations of effectors (Dreyer et al. 1999; Sander et al. 2003). Annotation of the complete 423 member human KRAB-KZNF locus set indicated that all three of these general types of alternative transcripts are common products of the genes in this family (Table 1; Suppl. Table S1). Transcripts encoding fingers-only variants are often generated by splicing events that skip KRAB-encoding exons, or by the use of alternate transcription start sites located downstream of effector exons. By contrast, KRAB-only variants are typically generated by inclusion of an additional exon between the KRAB and zinc finger exons, which throws the ZNF-encoding exon out of frame, or through the use of an alternate 3' exon, either splicing out the finger exon or stopping short of it (e.g., transcript model ZNF681(Aboobaker and Blaxter), Fig. 1).

Alternative splicing occurring within the large zinc finger-encoding exon was observed for 17 loci, resulting in a predicted protein with an altered ZNF array. By removing zinc-finger motifs, this type of alternative splicing event could generate protein isoforms with different DNA binding specificities. RNA species with structures that suggest splicing events within the finger arrays that remove one or two fingers but maintain the reading frame of the downstream finger motifs were also observed for several genes. However, we did not detect canonical splice signal sequences at these putative splice positions and the biological relevance of these mRNA species is therefore uncertain.

Altogether, we identified 818 different transcripts generated from the 423 KRAB-ZNF genes; together these transcripts encode a total of 743 distinct proteins. The most prevalent class of transcripts exhibits the archetypal KRAB-ZNF gene structure, with a KRAB-A, KRAB-B and ZNF-encoding exons (221 genes), but mRNA species encoding proteins with several other combinations of effector domains were also frequently found (Table 1). Predicted KRAB-ZNF proteins with 2-40 tandem zinc-fingers are encoded by the collection of human genes, with a median number of 12.5 ZNF motifs per gene. SCAN-containing proteins typically include a smaller number of zinc finger motifs than their KRAB-containing counterparts, with a median number of 8 ZNF motifs in SCAN-KRAB proteins and 5.5 ZNF motifs for proteins including the SCAN domain alone (Table 1). This observation suggests that the SCAN proteins might typically recognize shorter DNA binding motifs, an interpretation that is interesting in light of the fact that many SCAN-ZNF proteins have been shown to bind DNA as homodimers, potentially recognizing DNA sequences that are arranged as tandem repeats (Sander et al., 2002).

Gene fragments and pseudogenes

We identified 227 gene fragments and 39 full-length pseudogenes. Gene fragments were found frequently distributed among the protein-coding genes in KZNF clusters. Analysis of those pseudogenes included in a published set of recent human segmental duplications (Bailey et al. 2001) confirmed that most arose from neighboring, intact genes by partial-gene duplication events. However, in previous studies we have also documented gene remnants left behind after lineage-specific deletions (Hamilton et al. 2003; Shannon et al. 2003) and the histories of these gene fragments therefore cannot be ascertained without a more detailed phylogenetic analysis.

In addition to segmental duplicates, we also found evidence of 75 processed KRAB-ZNF pseudogene sequences. Three of these processed pseudogenes maintain open reading frames potentially capable of encoding functional proteins (LLNL1071 (HSA3, 32Mb), LLNL1040 (HSA9, 35Mb), and LLNL973 (HSA12, 132Mb)), but all other retroposed pseudogenes correspond to degraded, nonfunctional copies.

Gene Clustering and Evolution

As has been noted previously, most KRAB-ZNF genes reside in familial clusters (Bellefroid et al. 1995; Dehal et al. 2001; Shannon et al. 2003; Shannon and Stubbs 1998) Defining "clusters" liberally as 2 or more related neighboring loci separated by less than 200kb of intergenic sequence, we identified 63 KRAB-ZNF, SCAN-ZNF and mixed clusters in the human genome (Suppl. Table 1). A total of 382 intact KRAB-ZNF genes reside in these clusters; in addition, 41 KRAB-ZNF loci can be classified as isolated 'singleton' genes (i.e. not included in any cluster grouping) (Table 1; Fig. 2). Most of the genes are concentrated in 25 of the 61 human gene clusters, each of which contains 4 or more loci (Table 2).

To identify evolutionary relationships, we constructed a phylogenetic tree based on KRAB-A-encoding nucleotide sequences from the full set of predicted human KRAB-ZNF genes (Fig. 3). As suggested by previous studies, evolutionary relatedness is generally associated with physical proximity in this family. For instance, the tree defines clear associations of large sets of genes from the major chromosome 19 clusters located at 11-12 Mb (cluster 18 in Table 2; colored light green on Fig. 3), 19-23 Mb (cluster 19; orange), 41-42 Mb (cluster 21; pink), 57-58 Mb (cluster 23; blue-green), and 60-63 Mb (clusters 24 and 25; violet). Some intermixing of related genes from different chromosomal locations was also seen. For example, KRAB-C-containing genes, ZNF101 and ZNF14, located at the p-telomeric end of HSA19 cluster 19, are not related to neighboring cluster members but find their closest relatives in cluster 18, located 8 Mb away (Table 2). Three genes located in a mixed-family cluster at HSA1 are also closely related to this KRAB-C gene clade. All other members of the large centromeric HSA19 cluster group together in the tree with relatives residing in HSA7 and at several other distributed sites (Fig. 3). These data confirm that tandem *in situ* duplication events have represented the major mechanism of new gene creation in the KRAB-ZNF family, but also indicate that distributed duplication events have played a prominent role.

Most genes containing the KRAB-A motif also include the KRAB-B modulator, or less common modulators KRAB-b, KRAB-BL or KRAB-C (Table 1). These associations appear within separate clades in the KRAB-A-based tree, indicating that these motifs arose and were expanded within specific families (Fig. 3). Unlike the results for the KRAB modulators, the SCAN-containing KRAB-ZNF genes do not group together in one evolutionary clade (red circles, Fig. 3). This pattern could be explained if the SCAN-KRAB-ZNF combination is ancient, with a history of frequent loss of one or the other effector domains during the expansion of the gene family. This kind of history would also explain related genes with different combinations of SCAN and KRAB effector motifs observed in several clusters (Suppl. Table 1). Recent, lineage-specific expansions in gene number can affect both KRAB-ZNF genes (Hamilton et al. 2003) and SCAN-containing genes (Sander et al. 2003). It is also possible that the SCAN-KRAB combination arose more than once, since SCAN-KRAB-ZNF loci with different combinations of KRAB-A and KRAB-B domains are also distributed in several different clades.

Phylogenetic analyses also highlighted significant similarity between cluster members and isolated loci distributed at distant chromosomal sites. Some of these duplicates are part of larger segmental duplications that also include additional ZNF family members. For example, seven KRAB-ZNF genes and one pseudogene sequence distributed in HSA4, 8, 11, and 12 (the ‘scout clade’ highlighted in Fig. 3), show 96-98% nucleotide sequence identity over the lengths of duplication units spanning more than 70 kb. This degree of sequence similarity indicates that most of the duplication events occurred ≤ 15 Myr ago, and several gene copies may have arisen after the divergence of the hominid lineage. Interestingly, the recently evolved HSA8 copies from this gene family (LLNL1035, LLNL1101, LLNL1102, and LLNL1103; Suppl. Table 1 and http://znf.llnl.gov) are contained within segmental duplications that have been shown to vary in copy number in the genomes of individual humans (Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005).

Paralogs, orthologs and recent primate duplications

To identify putative orthologs of human KRAB-ZNF genes, we also generated HMMER-based gene models from the chimpanzee, mouse and dog genomes and searched for reciprocal best BLAST matches between these models and the curated human coding gene set (Suppl. Tables 2-4). The draft status of mouse, dog and chimp genomes makes it impossible to determine the KRAB-ZNF gene repertoires of those species completely, and we did not curate models from these species to distinguish genes from intact pseudogenes. However, these data provided a preliminary view of KRAB-ZNF gene conservation in mammals. We identified 108 sets of putative human-mouse orthologous pairs (defined as clear reciprocal best matches in BLAST; Suppl. Table 5). This conserved gene set includes only 76 of 365 KRAB-ZNF genes without a SCAN domain (21% of total), but 32 of the 58 total human SCAN-domain containing ZNF genes (55%) (Table 1). The relative enrichment of SCAN-containing loci in the conserved gene set suggests that SCAN-containing gene family members may be evolving under different constraints than their KRAB-only counterparts.

In addition, many human proteins matched more than one putative mouse or dog ortholog with high similarity, and conversely, examples were detected in which multiple human genes identified the same single reciprocal best-matching gene in dog or mouse (Suppl. Table 5). Three hundred forty-six of the 423 human proteins detected a clear chimpanzee ortholog, and all human KRAB-ZNF clusters with 4 or more genes are represented by related clusters with similar numbers of loci in the draft chimpanzee genome (Table 2). A total of 212 human KRAB-ZNF protein-coding loci residing in 34 clusters detected a believable homolog in either the mouse or the dog genomes. The remaining 211 loci represent potential primate-specific genes (Suppl. Table 6).

Included in this set of 211 loci are KRAB-ZNF genes residing in a centromere-adjacent HSA19 cluster (cluster 19, at 19-23 Mb), which has long been known to be primate-specific (Bellefroid et al. 1995). Genes in this cluster date back to early primate evolution, but the family underwent a significant expansion sometime before the divergence of old and new world monkeys (Eichler et al. 1998). Subsequent duplications have created new genes within the HSA19 cluster, new primate-specific clusters in HSA4 and HSA7, and singleton gene copies at several distributed sites (Hamilton et al. submitted). Altogether, this expanded subfamily comprises 62 of the 211 candidate primate-specific protein-coding genes.

Other members of the potential primate-specific gene group have also been involved in recent duplications, as indicated by overlaps between annotated KRAB-ZNF loci and recent segmental duplications in the human genome. The dataset generated by Bailey and colleagues (2001) includes all duplicated regions with >90% identity over >1kb. Within these domains we counted genes spanned by duplicated segments over most or all of the transcription unit length. Forty-eight of the 211 human genes without recognizable homologs in mouse or dog sequence are spanned by these recent duplications; 12 additional genes within duplicated regions are conserved in non-primate species and potentially represent “parental” copies for some of the primate-specific duplicates (not shown).

To examine older primate duplications, we also compared noncoding sequences from within each locus with BLAST, identifying all gene pairs with >85% noncoding sequence identity. A total of 111 genes were identified with noncoding sequence BLAST matches at this level to other genes in the human KRAB-ZNF set, indicating their involvement in duplication events that have taken place within the past 30-45 Myr (Suppl. Table 6).

The majority of these putative primate-specific genes reside in identified primate-specific clusters or in clusters that have undergone significant primate-specific expansions in gene numbers (Table 2; Suppl. Table S6). In addition, a number of recent duplications have created one or more new copies of genes that are otherwise well conserved (Hamilton et al. 2003; Shannon et al. 2003). We see evidence of lineage-specific expansion in other species as well, such as the expansion of genes in the mouse genome illustrated in the cluster shown in Fig. 4B. However, not all clusters exhibit evidence of rapid evolutionary change. For example, one cluster located in HSA5, is unusually well conserved, with most cluster members conserved as 1:1 orthologous pairs in mouse (Fig. 4A) and dog (not shown, see Table S6 for a complete list of orthologous pairs between human and chimpanzee, mouse, and dog proteins). By combining the non-coding sequence similarity, segmental duplication, and primate specific cluster data, we determined that of the 211 human KRAB-ZNF genes without a 1:1 ortholog in dog or mouse, 136 genes (32% of all human KRAB-ZNF genes) show compelling evidence for primate specificity.

KRAB-ZNF Gene Expression, Cluster Position, and Evolutionary History

Microarray expression data is publicly available for 335 of the 423 predicted human KRAB-ZNF genes, corresponding primarily to members of the Refseq or known gene sets (Su et al. 2004); Affymetrix (http://www.affymetrix.com). The actual number of genes for which reliable gene expression data exists is smaller, since many of the probes in publicly available microarray sets fall within sequence regions shared by multiple related genes (a total of 508 probe sets are associated with these 335 genes). By focusing on microarray data from relatively unique sets of probes corresponding to 260 genes (Suppl. Table S8) we were able to obtain preliminary answers regarding the evolution and function of KRAB-ZNF genes. Specifically, we were interested in addressing two questions. First, we sought to determine if, as in certain other clustered gene families (e.g. homeobox, olfactory receptor, immune receptor, and hemoglobin genes (Dryer 2000; Hardison 1998; Holland 1992; Hughes 2002)), KRAB-ZNF genes are co-expressed, or alternatively, (Gan et al. 2000) if tissue-specific expression is separately controlled and free to diverge. Secondly, we wanted to determine if evolutionary relatedness of paralogs predicts a corresponding relatedness in expression patterns. To address these questions, we used the Cluster software package (Eisen et al. 1998) to determine relationships between microarray expression patterns of KRAB-ZNF genes and to compare similarity in expression patterns with locations.

Although this analysis revealed that gene expression patterns for the KRAB-ZNF family are diverse, some clustering was observed. Most notably, expression data for 66 (25%) of the human KRAB-ZNF genes analyzed clustered with >40% correlation into one major group with highest levels of expression in bone marrow, spleen, thymus, lymph nodes and related immune cell types. A second group of 52 genes (20%) displayed the opposite pattern, i.e. significantly reduced expression levels in immune-related tissues and cells (Fig. 5, Suppl. Table S8). However, expression pattern similarities do not correlate with location in the genome; rather, groups of genes with most similar expression patterns were generally derived from different chromosomal regions. These data confirm more generally the patterns we have observed in in-depth studies of specific KRAB-ZNF clusters (Hamilton et al. Submitted; Shannon et al. 2003). Although recently evolved duplicate gene pairs sometimes display overlapping patterns of expression (Shannon et al. 2003), neighboring genes were only infrequently grouped as best matches on the basis of expression similarity (Fig. 5). The wide divergence of gene expression patterns amongst closely related genes indicates that clustered genes are not co-expressed, and suggests that promoters and other regulatory elements duplicated along with KRAB-ZNF coding sequences are diverging rapidly to give rise to genes with distinct patterns of tissue-specific expression.

DISCUSSION

We have identified and curated 423 human loci capable of encoding complete KRAB-ZNF proteins, including 91 novel loci. RNA evidence indicates that the vast majority of these models correspond to expressed human genes, many of which are alternatively spliced to encode predicted protein isoforms with potentially different functional properties. For example, inclusion of a KRAB-B domain has been shown to enhance repressor activity of KRAB-A proteins (Vissing et al. 1995), whereas the inclusion or exclusion of a SCAN domain may facilitate or abolish dimer formation, respectively (Williams et al. 1999). Alternative splicing of a SCAN-KRAB gene to exclude the KRAB-encoding exon might have an even more dramatic effect, potentially transforming the protein into a transcriptional activator. The prevalence of these alternative transcripts therefore predicts a vast array of KRAB-ZNF proteins with a wide range of gene regulatory roles. Although the KRAB-ZNF gene family includes a number of isolated segmental duplicates, most family members appear to have been created by tandem *in situ* duplication events yielding 63 human familial gene clusters. Since patterns of tissue-specific expression do not generally correlate with genome location, the clustered organization of KRAB-ZNF loci is probably a simple reflection of this evolutionary history rather than a sign of co-regulation or other type of functional linkage.

Based on comparisons between the curated human gene set and uncurated gene models from draft mouse, dog and chimpanzee genomes, we present a preliminary classification of human KRAB-ZNF genes based on their degree of conservation or lineage-specificity. A total of 212 human genes identify clearly orthologous sequences in either mouse or dog of which 99 genes can be grouped in obvious 1:1 orthologous relationships that have been deeply conserved throughout mammalian evolution. A small number of KRAB-ZNF clusters are deeply conserved as clusters in gene content and structure, but most human clusters harbor genes that have arisen through lineage-specific duplication events and several human gene clusters detect no obvious syntenically homologous counterparts in dog or mouse.

Data derived from interspecies comparisons, BLAST similarity in noncoding intronic sequences, and overlap with recent human segmental duplications (Bailey et al. 2001) identified 111 loci that are likely to represent primate-specific genes. Among these loci, 62 genes are considered members of the ZNF91 subfamily (Hamilton et al. Submitted) and most of the remainder are found in conserved gene clusters and represent lineage-specific duplicates of more ancient genes. For a total of 111 of these putatively primate-specific genes, we obtained evidence of relatively recent duplication, providing additional support for their primate specificity. Many genes within the known primate-specific clusters have been shown to predate the divergence of old world and new world monkeys (Bellefroid et al., 1995; Eichler et al., 1998) and these and other older primate duplicates cannot be detected reliably through sequence similarity measurements. Confirmation of the evolutionary histories of these genes must await the completion of sequences from new world monkey and prosimian genomes.

Most non-KRAB ZNFs and virtually all invertebrate *Krüppel*-type proteins carry short ZNF arrays, typically comprised of 3-4 zinc-finger motifs; by contrast, the DNA binding arrays of mammalian KRAB-ZNF proteins include a median of 12 tandem fingers. The large number of ZNF motifs in the predicted human proteins suggests a preference for relatively long and specific DNA binding sites, a conjecture that is supported by the known binding sites of KRAB-ZNF and other 'polydactyl' ZNF proteins (e.g., (Gebelein and Urrutia 2001; Schoenherr and Anderson 1995; Tanaka et al. 2002; Zheng et al. 2000). However, the correlation between ZNF array length and binding site length is probably not dictated by a simple formula; for certain proteins at least, only subsets of zinc fingers are required for DNA binding (e.g. *ZBRK1,* (Zheng et al. 2000)) and different fingers may be used to recognize targets of distinct sequence composition and types (e.g., ZAC, (Hoffmann et al. 2003); CTCF*,* (Ohlsson et al. 2001)). The potential ability of these polydactyl proteins to interact with multiple distinct DNA recognition sequences through different subsets of zinc fingers could potentially multiply the already substantial functional diversity of the KRAB-ZNF family.

What kinds of biological roles would fit the dynamic evolutionary history of KRAB-ZNF proteins? Recent studies of two genes from a rodent-specific cluster shed a particularly interesting light on this question. *Mus*-specific KRAB-ZNF gene duplicates *Rsl1* and *Rsl2* *(*regulator of sex-limited expression, 1 and 2) regulate gender-specific expression of distinct target genes involved in reproduction, and *Rsl* mutations with clear effects on gene expression and subtle phenotypic effects have been described (Krebs et al. 2003). The known functions of these genes raises the possibility that other clustered KRAB-ZNF genes may also serve to regulate gender-specific traits, a set of physiological properties known to be under intense evolutionary selection (e.g., (Pischedda and Chippindale 2005)). One primate-specific KRAB-ZNF protein (ZNF91) has been implicated in the regulation of immune-related genes many of which are also evolving rapidly, and available microarray data also indicate a prominent role for KRAB-ZNF genes in immune cell function.

However, since gene targets are known for only a handful of KRAB-ZNF proteins, the cumulative biological impact of their remarkable evolutionary history remains a matter of conjecture. The target genes that have been identified for both highly conserved (e.g. ZNF202, ZBRK1, KRIM1, NRIF, ZNF354c) and lineage specific KRAB-ZNF genes (RSL1 and 2; ZFP57; ZNF253), suggest intriguing roles for these transcriptional repressors in a wide variety of biological processes and pathways. Because of their sheer numbers, their wide range of tissue-specific expression, and their dynamic evolutionary history, we predict that KRAB-ZNF genes have played a significant role in shaping human biology, including both primate-specific and deeply conserved traits. Given the fact that this single gene family accounts for more than 20% of the 2000 predicted human transcription factor protein-coding loci (Messina et al. 2004), a more complete understanding of the functions of the KRAB-ZNF family will be essential not only for understanding pathways of vertebrate evolutionary diversity, but also for building accurate models of gene regulation and its role in human disease.

METHODS

Genome searches and initial data analysis

Existing Refseq human KRAB-A, KRAB-B, KRAB-b, KRAB-C, and SCAN protein sequences were collected and trimmed to include only motif residues completely encoded within single exons. All finger protein sequences (X7-C-X2-C-X12-H-X3-H) from HSA19 were collected by a simple pattern matching script. Sequence alignments for each motif-type were generated using CLUSTALX (Thompson et al. 1997) and submitted to the HMMER profile HMM matrix building tool HMMBUILD to generate matrices. These matrices, along with Pfam (Bateman et al. 2004) BTB matrix PF00651 were used by the HMMER search program to identify all putative motif matches in a six-frame translation sequence set of the Hg17 DNA sequences. In addition, DNA sequences from exons immediately preceding known KRAB-A exons were used to search Hg17 chromosomal sequences using BLAST (Altschul et al. 1990). Output from the HMMER and BLAST searches was compared to the genomic coordinates of publicly available gene models from Refseq (the NCBI mRNA reference sequence collection), UCSC Known (known protein-coding genes based on protein data from UniProt (SWISS-PROT and TrEMBL) and mRNA data from Refseq), and MGC (Mammalian Gene Collection gene models) (Strausberg et al. 1999) to identify previously characterized loci. The search results were also arranged in chromosomal order and grouped based on proximity and orientation into putative loci.

The human HMM matrices were also used to search the chimp (PanTro1), mouse (mm6) and dog (canFam1) genome translations, and putative loci were generated based on proximity and orientation as above. Crude protein sequences for these non-human loci were generated by extending from motif coordinates N- and C-terminally until a translational stop signal was encountered, eliminating overlapping sequences from adjacent motifs, and joining all collected sequences for each locus.

Gene annotation and database curating

All editing of gene model structures was performed with the sequence annotation editor APOLLO (http://www.fruitfly.org/annot/apollo) and existing publicly available annotation data (e.g., mRNA, EST, CpG islands, first exon prediction, gene prediction, etc.), as well as HMMER motifs generated in house. Public gene models were modified where transcript adjustment was warranted based on RNA evidence. New models were created based on RNA evidence where available, but in some cases, based solely on motif locations generating an open reading frame with appropriate splice boundaries within a locus. All intron boundaries were compared to the canonical splice sites (GT-AG, AT-AG, and GC-AG) and models were adjusted accordingly. UTRs of gene models were maximally extended when significant RNA evidence suggested an extension of an existing model's 5′ or 3′ ends was justified. Loci were considered to contain pseudogenes if no gene model could be made which could produce a functional protein. Loci producing KLF-like (2-4 fingers on separate exons), PRDM-like (>4 fingers on more than 2 exons) and BTB-containing genes were excluded from the ZNF dataset. Adjacent genes were considered "clustered" if the intergenic sequence separating the genes was less than 200 Kb.

Comparative genomic analyses

Protein sequences from all human KRAB and SCAN-KRAB loci were separately BLASTed against databases containing all human ZNF protein sequences or all crude sequences from chimp, mouse, or dog loci. The best match (non-self match when searching for paralogs) was then BLASTed against the human database. If the best match returned in this second search was the initial search query, the pair was considered a reciprocal pair. Data was collected for the best six matches for each protein, and reciprocal matches were flagged. Chromosomal positions of reciprocal matches were compared, and relative position comparisons were used to analyze evolution of loci and clusters both within the human genome and across the four species.

Evolutionary analysis

A tree of phylogenetic relationships was generated for the KRAB-A exons of 407 KRAB-ZNF loci. The KRAB-A was chosen for analysis because it is the most common effector domain in the family of KRAB-ZNF genes, and is easily aligned between the large number of loci. Intact KRAB-A sequences were extracted from the catalog loci and alignments of the sequences were made using CLUSTALX. The alignment was manually checked using SeAl (Rambaut 1996). The PAUP 4.0b10 package (Swofford 2002) was used to generate trees using mean character differences and the neighbor-joining (NJ) method. A Xenopus KRAB-A (Xfin) sequence was added as a potential outgroup. Genes that had a KRAB-A and also an in-frame, translated KRAB-b, KRAB-BL, KRAB-C or SCAN domain were marked on the phylogeny.

Selection of microarray probe sets and expression clustering

Affymetrix-based GNF Atlas expression data (Su et al. 2004) was obtained from the UCSC genome Bioinformatics website and GNF1H and U133A probe sets relevant to genes in our dataset were selected. Due to the possibility that probe sequences designed for recently-duplicated ZNF genes may significantly match sequences of paralogous genes, a screening process was designed to select the most unique probe set for each gene. First, the individual sequences for all probe sets from both chips were BLASTed against all catalog transcript sequences (including non-canonical genes). A pool of gene-specific probe sets was then identified based on two criteria: first, all sequences within a given probe set perfectly aligned to some portion of a single target locus; second, the ration of the number of target-specific alignments over the total number of alignments to any locus (i.e. target and non-target loci to which probes had ≥ 20/25 matches) was greater than 0.8. This second requirement eliminates probe sets which interrogate additional target(s) well or numerous targets poorly. In instances where multiple candidate probe sets were identified for a single gene, a representative probe set was manually chosen based on comparing where in the transcripts the probe sequences aligned.

Using Cluster v. 2.11 (Eisen et al. 1998), the expression profiles for the selected probe sets were hierarchically clustered using an uncentered correlation metric. The resulting cdt files were visualized using TreeView v. 1.60 (Eisen et al. 1998) to generate Figure (Expression (B)). Profiles were also manually arranged in chromosomal order and visualized in TreeView to generate Figure 5.

ACKNOWLEDGMENTS

We thank Ivan Ovcharenko for advice on programming and genome searching methods, and David Goodstein and Astrid Terry at the Joint Genome Institute for advice on Apollo and gene annotation. We also thank Colleen Elso, Jutta Kollet, Jason Raymond, and Alice Yamada for critical reviews of the manuscript and website.

This work was performed under the auspices of the U. S. Department of Energy (DOE) by the University of California, Lawrence Livermore National Laboratory (LLNL) under Contract No. W-7405-Eng-48. The project (04-ERD-084) was funded by the Laboratory Directed Research and Development Program at LLNL.

TABLE AND FIGURE LEGENDS.

Table 1. Characteristics of KRAB / SCAN-KRAB ZNF gene families.

Table 2. Overview of major human gene clusters and relative conservation in other mammals.

Figure 1. Schematic of locus ZNF681 adapted from the UCSC Genome Bioinformatics Genome Browser with ZNF data added as additional tracks. Curated gene model predictions are shown at the top of the image (green); bracketed numbers indicate unique catalog model identifiers. All Refseq and UCSC Known gene models for locus ZNF681 are shown at the bottom of the image (black). Protein motifs identified in the genomic sequence are shown as labeled in the center, with sequences containing stop codons (purple) separated from "intact" motifs (blue). All gene models are represented in the image as follows: thin bars represent untranslated exon sequence, thick bars are translated exon sequences, and arrowed lines represent intronic sequence and direction of transcription.

Figure 2. Genome-wide distribution of KRAB / SCAN-KRAB genes. Positions of catalog genes are shown in relationship to each chromosome. Gray bars represent the total chromosomal sequence, with light gray (white?) segments representing telomeres, centromeres, and gaps in the chromosome sequence. In the whole genome map (Panel A), individual loci or clusters of loci are represented by green bars, with bar width indicating the relative span the loci or clusters of loci include relative to the chromosome. A map of chromosome 19 (Panel B) shows more detail of the loci relative to the chromosome. Light green bars represent clusters of genes; dark green bars are individual genes. Blue and red bars represent pseudogenes / fragments and retroposed pseudogenes, respectively. Maps of all chromosomes are available at the catalog website (http://znf.llnl.gov).

Figure 3. Phylogenetic tree of human KRAB-A motifs.

A neighbor-joining phylogenetic tree including 407 human KRAB-A sequences. Gene designations are removed from this unrooted phylogram for clarity but genes from several major physical clusters are colored to show correspondence between location and sequence similarity. White or differently-colored circles within a clade surrounded by a larger block of color represent genes that are related to those from a particular cluster. KRAB-A motifs associated with a SCAN domain are marked with red circles. KRAB-B motifs were not highlighted in order to distinguish the KRAB-A's with less common modulators, indicated as described in the figure. The green arrow notes the position of the Xenopus KRAB sequence Xfin, added as a potential outgroup. In several instances, such as within the 'scout clade', KRAB sequences were found to be identical between closely related genes.

Figure 4. Comparison of conserved clusters of KRAB genes. Groups of human genes defined to be clustered (see Methods) were compared to genes from other species and adjacent non-ZNF genes were identified. The relative genomic locations of the genes (HSA, human; PT, chimpanzee; and Mmu, mouse) were mapped as shown in the figures (colored arrowheads), with orthologs indicated by vertical lines and similar gene relationships indicated by dashed lines. Non-ZNF genes are colored gray, KRAB genes are colored black, and other ZNF genes are colored white. In the cluster shown in panel A, most genes are well conserved, with only one possible expansion of a non-KRAB ZNF in chimpanzee. Panel B represents a cluster with little change between human and chimp, but a significant expansion of one human ortholog in mouse.

Figure 5. Analysis of expression patterns for KRAB and SCAN-KRAB ZNF genes. Expression levels of members of the human KRAB and SCAN-KRAB genes are represented by red (higher) and green (lower) boxes. Vertical columns of boxes represent different genes, and horizontal rows represent different tissues. We have grouped tissues into the following categories as labeled on the left of each panel: N, neural; I, immune; G, glandular; M, muscle; O, other organ; and R, reproductive. In panel A, the 260 analyzed genes are arranged in chromosomal order. Selected genomic clusters are indicated by colored boxes beneath the expression pattern. From left to right, the clusters are (as numbered in Table 2): 13, 14, 17, 18,

21, 22, 23, 24, 25, 6, 8, 9, and 10. In panel B, all selected expression profiles (described in text) have been clustered based on expression pattern similarity. Colored hash marks above the expression profiles indicate the chromosomal cluster from panel A with which each profile corresponds.

Table 1.

Gene type	No. found	No. in clusters	No. of singletons	Effector-only transcripts	ZNF-only transcripts	Median no. fingers	No. conserved in chimpanzee	No. conserved in mouse	No. conserved in dog
KRAB-A	91	81	10	9	27	12	71	20	36
KRAB-A-B	221	197	24	38	39	12	175	54	118
KRAB-A-BL	11	11	0	2	3	15	9	1	0
KRAB-A-b	11	11	0	1	3	15	11	0	2
KRAB-A-C	31	31	0	1	3	14	25	1	2
SCAN-KRAB-A*	28	23	5	6	0	8	26	17	23
SCAN	30	28	2	9	4	5.5	29	15	24
Total	423	382	41	66	79	12	346	108	205

* Includes all SCAN-KRABA genes with and without modulating motifs.

Table 2.

Human cluster		Conservation In Other Species[1]		
Cluster Number	Coordinates (Mb)	Chimpanzee	Dog	Mouse
1	chr1:243.3-243.8	+ +	+ -	+ -
2	chr3:40.49-40.55	+ +	+ +	- -
3	chr3:44.45-44.75	+ -	+ -	+ -
4	chr4:0.043-0.482	+ -	- -	- -
5	chr5:178.07-178.44	+ +	+ +	+ +
6	chr6:28.15-28.66	+ +	+ +	+ -
7	chr7:62.78-63.91	+ -	- -	- -
8	chr7:98.71-98.87	+ +	+ +	+ +
9	chr7:148.19-149.00	+ +	+ +	+ -
10	chr8:145.89-146.24	+ +	+ +	+ -
11	chr10:38.12-38.45	+ -	+ +	+ -
12	chr12:132.10-132.39	+ +	+ -	+ -
13	chr16:3.07-3.43	+ +	+ -	+ -
14	chr16:30.31-30.70	+ +	+ +	+ -
15	chr18:31.07-31.21	+ +	+ +	+ -
16	chr19:2.77-2.89	+ +	+ -	- -
17	chr19:9.11-9.75	+ +	+ -	+ -
18	chr19:11.56-12.60	+ -	+ -	+ -
19	chr19:19.63-24.11	+ -	- -	- -
20	chr19:39.72-40.14	+ +	+ -	- -
21	chr19:41.46-43.01	+ +	+ +	+ -
22	chr19:49.02-49.69	+ +	+ -	+ -
23	chr19:57.03-58.77	+ +	+ -	+ -
24	chr19:61.29-62.04	+ +	+ +	+ -
25	chr19:62.31-63.77	+ +	+ -	+ -

[1] Based on reciprocal BLAST results. Level of cluster conservation with each species indicated as follows: + +, at least ¾ of human genes conserved; + -, conservation, but less than ¾ of human loci conserved; - -, no conservation of human genes detected.

Figure 1.

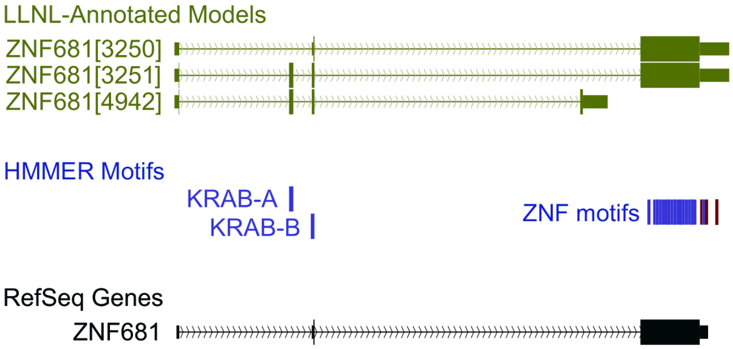

LLNL-Annotated Models
ZNF681[3250]
ZNF681[3251]
ZNF681[4942]
HMMER Motifs
KRAB-A
KRAB-B
ZNF motifs
RefSeq Genes
ZNF681

Figure 2.
(A)

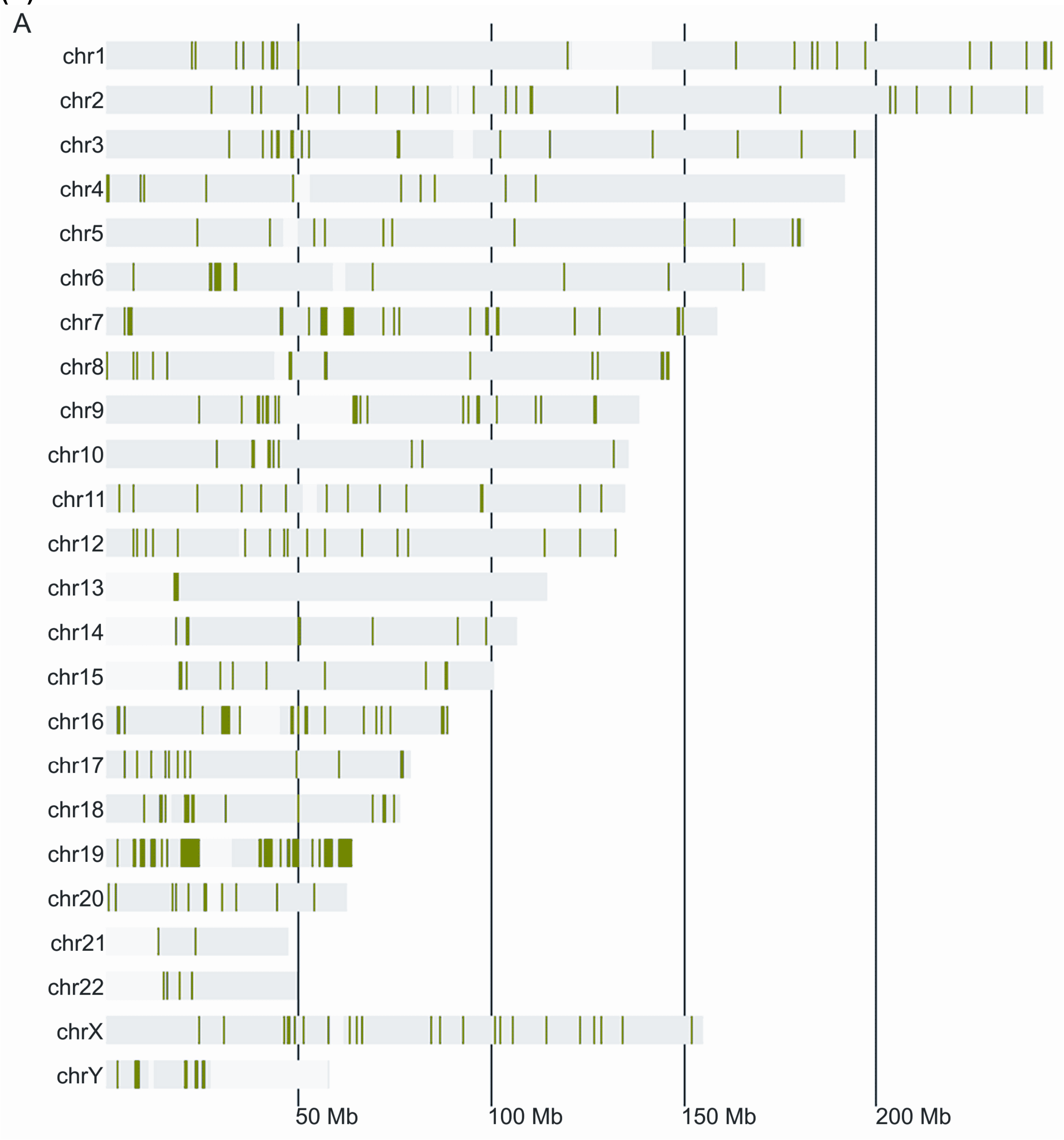

A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
50 Mb
100 Mb
150 Mb
200 Mb

Figure 2.
(B)

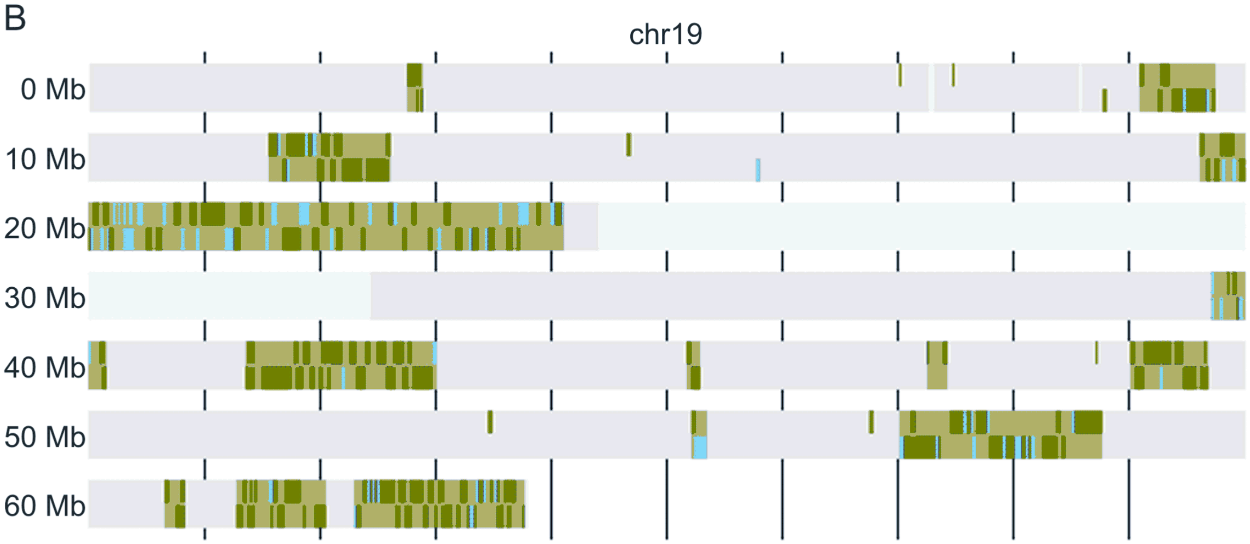


Figure 3.

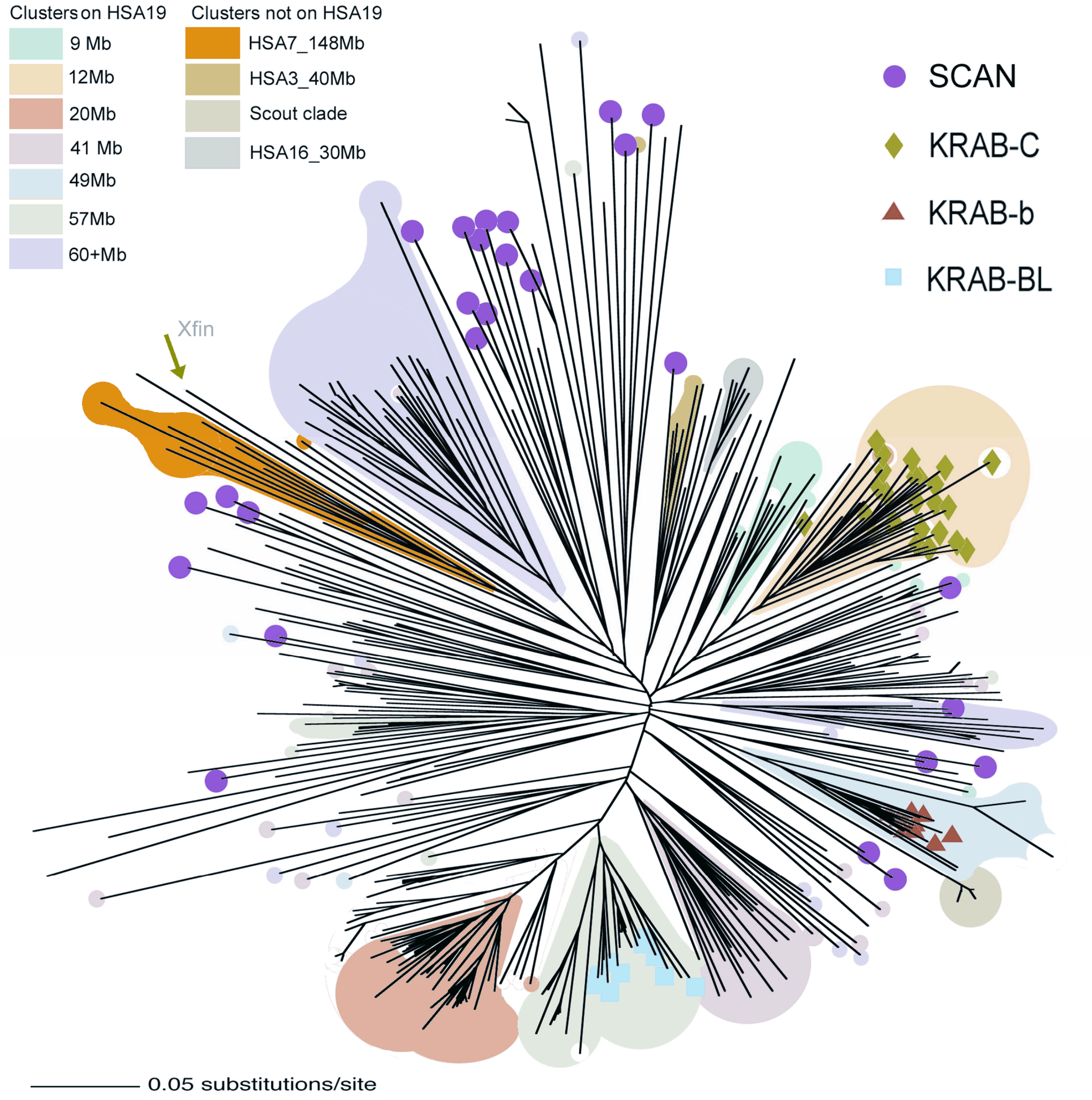

Clusters on HSA19
9 Mb
12Mb
20Mb
41 Mb
49Mb
57Mb
60+Mb
Clusters not on HSA19
HSA7_148Mb
HSA3_40Mb
Scout clade
HSA16_30Mb
SCAN
KRAB-C
KRAB-b
KRAB-BL
Xfin
0.05 substitutions/site

Figure 4.

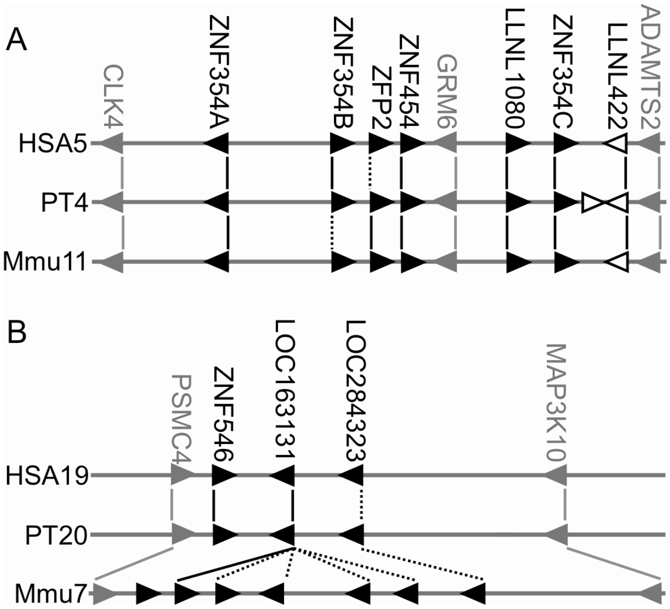
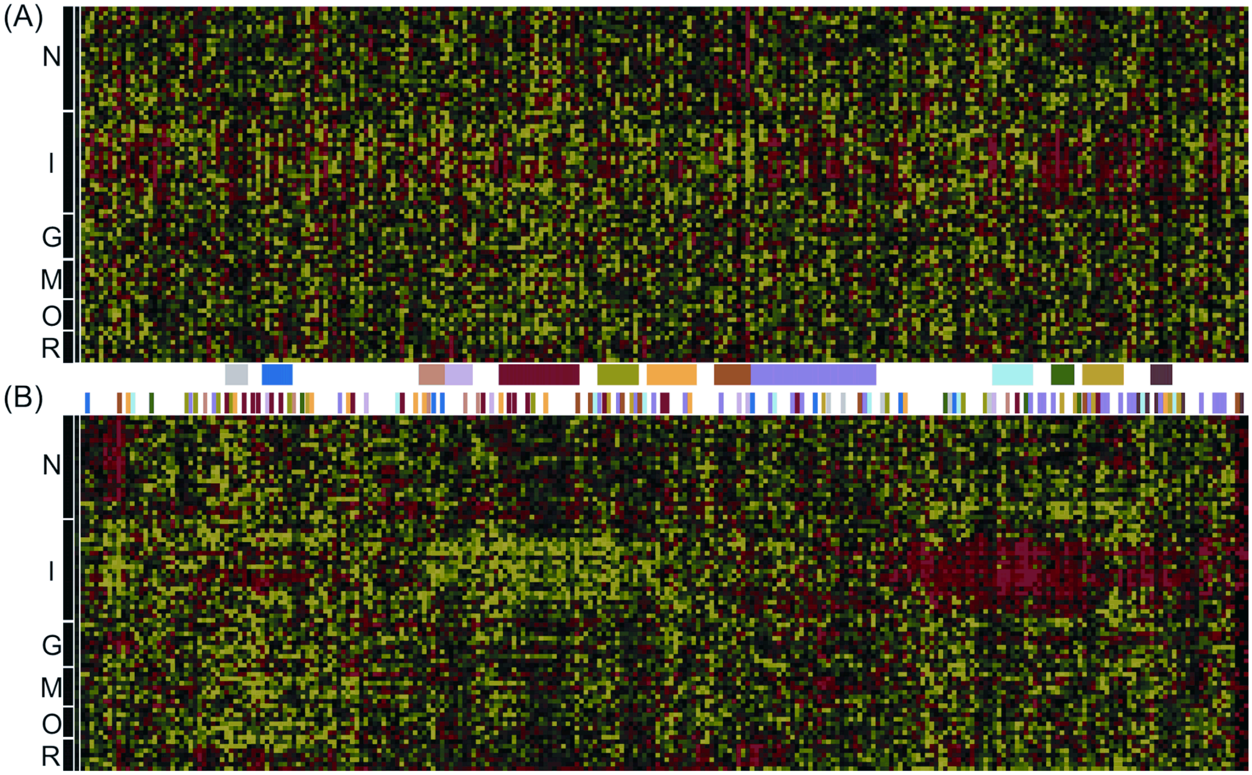

A
CLK4
ZNF354A
ZNF354B
ZFP2
ZNF454
GRM6
LLNL1080
ZNF354C
LLNL422
ADAMTS2
HSA5
PT4
Mmu11
B
PSMC4
ZNF546
LOC163131
LOC284323
MAP3K10
HSA19
PT20
Mmu7

Figure 5.



REFERENCES

Aboobaker, A.A. and M.L. Blaxter. 2003. Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. *Curr Biol* **13:** 37-40.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Ayyanathan, K., M.S. Lechner, P. Bell, G.G. Maul, D.C. Schultz, Y. Yamada, K. Tanaka, K. Torigoe, and F.J. Rauscher, 3rd. 2003. Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. *Genes Dev* **17:** 1855-1869.

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005-1017.

Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer et al. 2004. The Pfam protein families database. *Nucleic Acids Res* **32:** D138-141.

Bellefroid, E.J., J.C. Marine, A.G. Matera, C. Bourguignon, T. Desai, K.C. Healy, P. Bray-Ward, J.A. Martial, J.N. Ihle, and D.C. Ward. 1995. Emergence of the ZNF91 Kruppel-associated box-containing zinc finger gene family in the last common ancestor of anthropoidea. *Proc Natl Acad Sci U S A* **92:** 10757-10761.

Bellefroid, E.J., J.C. Marine, T. Ried, P.J. Lecocq, M. Riviere, C. Amemiya, D.A. Poncelet, P.G. Coulie, P. de Jong, C. Szpirer et al. 1993. Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. *Embo J* **12:** 1363-1374.

Bellefroid, E.J., D.A. Poncelet, P.J. Lecocq, O. Revelant, and J.A. Martial. 1991. The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proc Natl Acad Sci U S A* **88:** 3608-3612.

Chung, H.R., U. Schafer, H. Jackle, and S. Bohm. 2002. Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in *Drosophila*. *EMBO Rep* **3:** 1158-1162.

Collins, T., J.R. Stone, and A.J. Williams. 2001. All in the family: the BTB/POZ, KRAB, and SCAN domains. *Mol Cell Biol* **21:** 3609-3615.

Dehal, P., P. Predki, A.S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C.L. Ecale Zhou, S. Rash et al. 2001. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-111.

Dreyer, S.D., Q. Zheng, B. Zabel, A. Winterpacht, and B. Lee. 1999. Isolation, characterization, and mapping of a zinc finger gene, ZFP95, containing both a SCAN box and an alternatively spliced KRAB A domain. *Genomics* **62:** 119-122.

Dryer, L. 2000. Evolution of odorant receptors. *Bioessays* **22:** 803-810.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95:** 14863-14868.

Friedman, J.R., W.J. Fredericks, D.E. Jensen, D.W. Speicher, X.P. Huang, E.G. Neilson, and F.J. Rauscher, 3rd. 1996. KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes Dev* **10:** 2067-2078.

Gan, L., I. Lee, R. Smith, R. Argonza-Barrett, H. Lei, J. McCuaig, P. Moss, B. Paeper, and K. Wang. 2000. Sequencing and expression analysis of the serine protease gene cluster located in chromosome 19q13 region. *Gene* **257:** 119-130.

Gebelein, B. and R. Urrutia. 2001. Sequence-specific transcriptional repression by KS1, a multiple-zinc-finger-Kruppel-associated box protein. *Mol Cell Biol* **21:** 928-939.

Hamilton, A.T., S. Huntley, J. Kim, E. Branscomb, and L. Stubbs. 2003. Lineage-specific expansion of KRAB zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. *Cold Spring Harb Symp Quant Biol* **68:** 131-140.

Hamilton, A.T., S. Huntley, M. Tran-Gyamfi, D. Baggott, L. Gordon, and L. Stubbs. Submitted. Evolutionary expansion and divergence in a family of primate-specific zinc finger genes. *Genome Res*.

Hardison, R. 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J Exp Biol* **201 (Pt 8):** 1099-1117.

Hoffmann, A., E. Ciani, J. Boeckardt, F. Holsboer, L. Journot, and D. Spengler. 2003. Transcriptional activities of the zinc finger protein Zac are differentially controlled by DNA binding. *Mol Cell Biol* **23:** 988-1003.

Holland, P.W.H. 1992. Homeobox genes in vertebrate evolution. *BioEssays* **14:** 267-273.

Hughes, A.L. 2002. Natural selection and the diversification of vertebrate immune effectors. *Immunol Rev* **190:** 161-168.

Huntley, S., A. Hamilton, J. Kim, E. Branscomb, and L. Stubbs. In press. Tandem gene family expansion and genomic diversity. In *Comparative genomics: a guide to the analysis of eukaryotic genomes* (ed. M.D. Adams). Humana Press, NY.

Kim, J., A. Bergmann, S. Lucas, R. Stone, and L. Stubbs. 2004. Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. *Genomics* **84:** 47-58.

Knochel, W., A. Poting, M. Koster, T. el Baradi, W. Nietfeld, T. Bouwmeester, and T. Pieler. 1989. Evolutionary conserved modules associated with zinc fingers in *Xenopus laevis*. *Proc Natl Acad Sci U S A* **86:** 6097-6100.

Krebs, C.J., L.K. Larkins, S.M. Khan, and D.M. Robins. 2005. Expansion and diversification of KRAB zinc-finger genes within a cluster including Regulator of sex-limitation 1 and 2. *Genomics* **85:** 752-761.

Krebs, C.J., L.K. Larkins, R. Price, K.M. Tullis, R.D. Miller, and D.M. Robins. 2003. Regulator of sex-limitation (Rsl) encodes a pair of KRAB zinc-finger genes that control sexually dimorphic liver gene expression. *Genes Dev* **17:** 2664-2674.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Looman, C., M. Abrink, C. Mark, and L. Hellman. 2002. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol Biol Evol* **19:** 2118-2130.

Looman, C., L. Hellman, and M. Abrink. 2004. A novel Kruppel-Associated Box identified in a panel of mammalian zinc finger proteins. *Mamm Genome* **15:** 35-40.

Margolin, J.F., J.R. Friedman, W.K. Meyer, H. Vissing, H.J. Thiesen, and F.J. Rauscher, 3rd. 1994. Kruppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci U S A* **91:** 4509-4513.

Mark, C., M. Abrink, and L. Hellman. 1999. Comparative analysis of KRAB zinc finger proteins in rodents and man: evidence for several evolutionarily distinct subfamilies of KRAB zinc finger genes. *DNA Cell Biol* **18:** 381-396.

Messina, D.N., J. Glasscock, W. Gish, and M. Lovett. 2004. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res* **14:** 2041-2047.

Oh, H.J., Y. Li, and Y.F. Lau. 2005. Sry associates with the heterochromatin protein 1 complex by interacting with a KRAB domain protein. *Biol Reprod* **72:** 407-415.

Ohlsson, R., R. Renkawitz, and V. Lobanenkov. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17:** 520-527.

Pischedda, A. and A. Chippindale. 2005. Sex, mutation and fitness: asymmetric costs and routes to recovery through compensatory evolution. *J Evol Biol* **18:** 1115-1122.

Rambaut, A. 1996. Se-Al: Sequence alignment editor.

Sander, T.L., K.F. Stringer, J.L. Maki, P. Szauter, J.R. Stone, and T. Collins. 2003. The SCAN domain defines a large family of zinc finger transcription factors. *Gene* **310:** 29-38.

Schmidt, D. and R. Durrett. 2004. Adaptive evolution drives the diversification of zinc-finger binding domains. *Mol Biol Evol* **21:** 2326-2339.

Schoenherr, C.J. and D.J. Anderson. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267:** 1360-1363.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525-528.

Shannon, M., A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. 2003. Differential expansion of zinc-finger transcription factor Loci in homologous human and mouse gene clusters. *Genome Res* **13:** 1097-1110.

Shannon, M. and L. Stubbs. 1998. Analysis of homologous XRCC1-linked zinc-finger gene families in human and mouse: evidence for orthologous genes. *Genomics* **49:** 112-121.

Sharp, A.J., D.P. Locke, S.D. McGrath, Z. Cheng, J.A. Bailey, R.U. Vallente, L.M. Pertz, R.A. Clark, S. Schwartz, R. Segraves et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77:** 78-88.

Strausberg, R.L., E.A. Feingold, R.D. Klausner, and F.S. Collins. 1999. The mammalian gene collection. *Science* **286:** 455-457.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Swofford, D.L. 2002. PAUP* Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA.

Tanaka, K., N. Tsumaki, C.A. Kozak, Y. Matsumoto, F. Nakatani, Y. Iwamoto, and Y. Yamada. 2002. A Kruppel-associated box-zinc finger protein, NT2, represses cell-type-specific promoter activity of the alpha 2(XI) collagen gene. *Mol Cell Biol* **22:** 4256-4267.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727-732.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vissing, H., W.K. Meyer, L. Aagaard, N. Tommerup, and H.J. Thiesen. 1995. Repression of transcriptional activity by heterologous KRAB domains present in zinc finger proteins. *FEBS Lett* **369:** 153-157.

Williams, A.J., S.C. Blacklow, and T. Collins. 1999. The zinc finger-associated SCAN box is a conserved oligomerization domain. *Mol Cell Biol* **19:** 8526-8535.

Wu, Y., L. Yu, G. Bi, K. Luo, G. Zhou, and S. Zhao. 2003. Identification and characterization of two novel human SCAN domain-containing zinc finger genes ZNF396 and ZNF397. *Gene* **310:** 193-201.

Zheng, L., H. Pan, S. Li, A. Flesken-Nikitin, P.L. Chen, T.G. Boyer, and W.H. Lee. 2000. Sequence-specific transcriptional corepressor function for BRCA1 through a novel zinc finger protein, ZBRK1. *Mol Cell* **6:** 757-768.

WEB SITE REFERENCES

http://znf.llnl.gov, the ZNF database home page

http://hmmer.wustl.edu, HMMER tool web site

http://pfam.wustl.edu, the Pfam database of protein families and HMMs

http://www.fruitfly.org/annot/apollo, Apollo genome annotation tool website

http://www.affymetrix.com, Affymetrix human array probe data